Agent-Conditioned Multi-Contrast MRI Super-Resolution for Cross-Subject

Xinrong Hu✉
Wuhan Textile University
Wuhan, Hubei, China
hxr@wtu.edu.cn

Chao Fang
Wuhan Textile University
Wuhan, Hubei, China
18607259869@163.com

Yu Chen
Wuhan Textile University
Wuhan, Hubei, China
2215063028@mail.wtu.edu.cn

Kai Yang
Wuhan Textile University
Wuhan, Hubei, China
kyang@wtu.edu.cn

Chun-Mei Feng✉
Agency for Science, Technology and Research (A*STAR)
Singapore
strawberry.feng0304@gmail.com

Ping Li
The Hong Kong Polytechnic University
Hong Kong, China
p.li@polyu.edu.hk

Abstract

Multi-contrast MRI super-resolution (SR) techniques require the simultaneous acquisition of multiple contrasts from the same subject, which is often challenging in real-world clinical settings. In this paper, we propose a novel agent-conditioned multi-contrast MRI SR with cross-subject adaptation, termed AgentMRI. AgentMRI is the first attempt to improve the quality of target contrast images using external auxiliary contrasts from different subjects. It expands the traditional attention mechanism from a triplet to a quadruplet format (Query, Agent, Key, Value), where the agent can be trained to capture commonalities from the auxiliary contrast. These commonalities represent foundational anatomical and tissue structure features that are shareable, rather than details specific to a particular contrast. By interacting the agent with the target contrast, AgentMRI dynamically adjusts the model adapting the agent's knowledge to the target contrast image. This adapting process assists in identifying inherent connections between the auxiliary and target contrasts, even when they are not directly paired. Our extensive testing on fastMRI and clinical datasets demonstrates that our AgentMRI sets a new benchmark, surpassing state-of-the-art methods across various evaluation metrics.

Keywords: ***Magnetic resonance imaging, super-resolution, agent attention.***

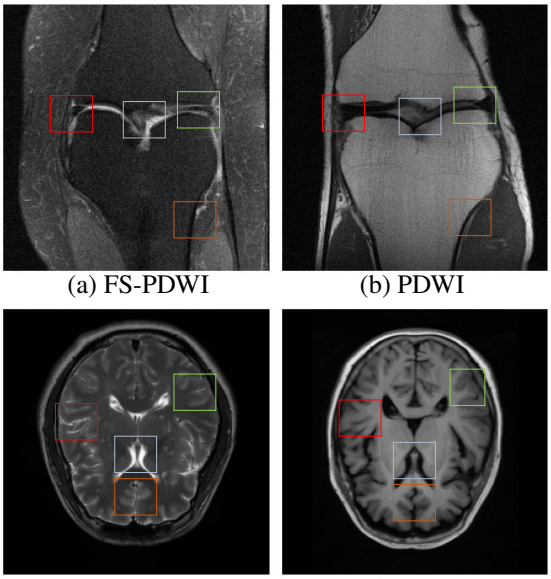
1. Introduction

Magnetic Resonance Imaging (MRI) is a non-invasive technique that generates images of the human body's internal tissues using strong magnetic fields and radiofrequency pulses [11]. Compared to other medical imaging technologies, such as Computed Tomography (CT) and Positron Emission Tomography (PET), MRI offers several advantages, including superior soft tissue contrast, absence of ionizing radiation, and the ability to acquire functional and metabolic information. This makes MRI particularly useful in neurological, musculoskeletal, cardiovascular, and oncological imaging. However, MRI also has limitations that impact its clinical utility. One primary limitation is its longer imaging times, which can lead to patient discomfort, increased susceptibility to motion artifacts, and higher costs due to prolonged use of the MRI system [32]. Another limitation is the lower signal-to-noise ratio (SNR) in MRI compared to CT and PET. Factors like magnetic field strength, coil quality, and imaging parameters affect SNR. Lower SNR can result in grainier images, obscuring fine anatomical details crucial for accurate diagnosis and treatment [41]. Therefore, accelerating MR imaging and improving its SNR have become prominent research topics.

In clinical settings, MR scanners sequentially acquire images in different modalities following specific imaging protocols tailored to diagnostic needs. For the same subject, each imaging modality often provides consistent data across multi-contrast and modality-specific anatomical and physiological information. The multi-contrast attributes of MRI have inspired researchers to exploit the complementary information among modalities that share analogous

anatomical structures [6, 8, 30], where T1-weighted images (T1WIs) and T2-weighted images (T2WIs), as well as proton density and fat-suppressed proton density-weighted images (PDWIs and FS-PDWIs), provide complementary structural perspectives. Due to the inherent physical properties of MRI, T1WI is generally easier to acquire than T2WI because it requires shorter repetition times (TR) and echo times (TE). Specifically, within the same imaging sequence, the acquisition time for T2WI tends to exceed that for T1WI due to the longer TR requirements of T2, as seen with T2SE acquisition times being longer than those for T1SE. This multi-modal imaging strategy utilizes faster-acquired modalities as auxiliary contrast images to guide and accelerate the imaging of target modalities with slower acquisition speeds, which has been verified by previous study [23,25,31]. For example, bicubic interpolation, Lanczos resampling, sparse representation, dictionary learning, and patch-based methods have been extensively applied in multi-modal MR imaging SR tasks. Qu *et al*. [35] integrated nonlocal means filtering with parallel imaging for MRI SR. Wang *et al*. [44] utilized Laplacian pyramids and adaptive sparse representation for multi-modal medical image fusion to achieve MRI SR work. Recently, the focus has shifted towards leveraging the capabilities of deep learning to address the challenges inherent in MRI SR. For example, Feng *et al*. [6] designed a multi-stage integration network that explores the dependencies among hierarchical stages in multi-contrast images. Li *et al*. [25] used transformer attributes to transfer the contextual information from auxiliary contrast to target contrast features across different scales, which significantly enhances the super-resolution quality of images. Li *et al*. [28] introduced a novel diffusion model for multi-contrast MRI SR which fully utilizes the prior knowledge from the diffusion model (DM) to ensure that the reconstructed MR images remain undistorted.

Despite significant advancements in processing multi-contrast MRI data, practical applications still face challenges in acquiring complete datasets of all contrasts for each patient due to regional development disparities and acquisition time and cost constraints [10]. Indeed, we observe that unpaired multi-contrast data still share commonalities, such as fundamental anatomical structure features (e.g., joint spaces, cartilage areas, cerebral cortex regions, and ventricular systems) present in the images. As shown in Fig. 1, (a) and (b) are a pair of FS-PDW and PDW knee MR images from different subjects in the fastMRI dataset, and (c) and (d) are a pair of T2W and T1W brain MRI images from different subjects in a real-world clinical dataset. Although these MR images are from different individuals, certain consistent anatomical and tissue structural features are present across different modalities and subjects. These common features offer constructive guidance for the SR of target images. Moreover, learning specific knowledge between contrasts, such as the capability of T1WI to describe morphological and structural information, can effectively complement the generation of T2WI [1,7,27,51]. This complementarity of multi-contrast information encourages us to explore multi-contrast features in MRI SR tasks, thereby enhancing image quality and diagnostic precision. Importantly, the vast online resources of medical images remain underutilized, motivating us to explore the correlations between different contrasts as well as subjects and repurpose extensive publicly available medical image data to generate high-quality target images from unpaired data [2].



(a) FS-PDWI (b) PDWI (c) T2WI (d) T1WI

Figure 1. Examples of four different imaging modalities. Images (a) and (b) show fat-suppressed proton density-weighted imaging (FS-PDWI) and proton density-weighted imaging (PDWI) from different subjects in the fastMRI dataset. Images (c) and (d) represent T2-weighted imaging (T2WI) and T1-weighted imaging (T1WI) from different subjects in a real-world clinical dataset. Despite their differences (marked by the red boxes), these different modalities from different subjects share fundamental anatomical and tissue structure features (marked by the green, blue-gray, and orange boxes).

To tackle this challenge, we introduce AgentMRI, an agent-conditioned technique for multi-contrast MRI SR that embraces cross-subject adaptability. This approach is designed to uncover both the *commonalities* and *distinctiveness* present in data across subjects, ultimately enhancing the target contrast. Different from previous methods that rely on paired images, despite the absence of paired images, AgentMRI expands the traditional attention mechanism from a triplet to a quadruplet format (Query, ***Agent***, Key, Value) to extract and capture *shareable foundational anatomical and tissue structure features* from auxiliary contrast. The interaction between the agent and the target con-

trast guides the model to focus on task-specific features, adapting auxiliary contrasts in the cross-subject to the target image. This adaptation process helps to identify the intrinsic connections between auxiliary and target contrasts, even when they are not directly paired.

For clarity, the main contributions of our work are summarized as follows:

- AgentMRI is the *first* to achieve SR of target contrasts using cross-subject auxiliary contrast images. Clinically, AgentMRI enables cost-effective multi-contrast MRI using public datasets for auxiliary contrast without requiring pairing with target images.
- We develop an *agent-conditioned* multi-contrast MRI SR approach that employs agent vectors to enable the model to transfer knowledge across *different contrasts and subjects*. This method adapts cross-subject auxiliary contrast data through an agent-conditioned mechanism, infusing beneficial features into the target image, e.g., foundational anatomical structures, and supplementary information.
- We conducted extensive experiments on the fastMRI and clinical medical image datasets, demonstrating that AgentMRI yields superior results over the state-of-the-art.

2. Related Work

2.1. MR Imaging

In clinical practice, magnetic resonance imaging (MRI) provides excellent soft-tissue contrast for clinical diagnosis and research. Image SR and reconstruction significantly improve the quality and speed of MRI imaging. Traditional MR image SR methods, such as bicubic interpolation [4], iterative deblurring algorithms [16, 39], and compressed sensing (CS) [12], have made significant strides in multi-frame MR image SR tasks. However, these methods offer limited information when processing individual images and usually rely on prior data information. In recent years, deep learning-based SR methods have demonstrated superior performance due to their ability to fully exploit the inherent attributes of images contained in extensive training datasets. For instance, Qui *et al*. [34] applied convolutional neural networks (CNNs) for knee MR image SR, and Lyu *et al*. [30] used ensemble learning for brain MR image SR. Jin *et al*. [20] employed UNet to capture spatial information for addressing inverse problems in MRI. More recently, generative adversarial networks (GANs) have been applied to MR images SR. For example, Jiang *et al*. [18] proposed a fused attentive generative adversarial networks framework to generate SR MR images from low-resolution (LR) MR images. Li *et al*. [26] incorporated attention mechanisms and cyclic loss within GANs for pelvic image SR. To learn essential regional feature representations from single MR images, Zhang *et al*. [49] introduced the squeezed and inspired inference attention network, demonstrating its effectiveness. Qui *et al*. [33] developed a gradual back-projection residual attention network to reconstruct MR images SR. Feng *et al*. [9] proposed an end-to-end task transformer network that integrates MRI reconstruction and SR into a single framework. Salvetti *et al*. [37] introduced a residual attention model using 3D convolutions and nested residual connections for multi-image super-resolution in remote sensing. However, the above methods usually focus only on single-contrast images, such as T1WI or T2WI, ignoring the multi-contrast information in MRI data.

2.2. Multi-modal Representation Learning

Multi-modal representation learning [19] extracts shared features from diverse data modalities such as text, images, and speech. For example, Kwon *et al*. [22] used the inherent properties of image-text paired data to implicitly learn the cross-modal alignment between language tokens and image patches for reconstructing the masked signal of one modality using another. Liu *et al*. [29] proposed an autoencoder-based multi-view missing data completion framework to learn a general representation of Alzheimer's diagnosis. Given the strong capability of multi-modal technology in representation learning [47], it has recently been widely applied in medical images. For instance, Tsai *et al*. [40] proposed a multi-modal transformer for unaligned multi-modal MRI sequences, demonstrating the effective integration of heterogeneous MRI data. Zhang *et al*. [48] utilized graph neural networks to fuse PET and MRI data, enhancing tumor segmentation and treatment planning by using spatial and functional information from both modalities. In contrast to the multi-contrast MR image segmentation task, the super-resolution (SR) task divides images into auxiliary and target contrasts. Given its shorter acquisition time, the auxiliary contrast can guide the restoration of the target contrast. For example, Lyu *et al*. [30] demonstrated that fusing multi-contrast information in high-level feature spaces yields superior results compared to low-level pixel-based combinations. Li *et al*. [25] pioneered hierarchical transformer networks for joint multi-contrast MRI reconstruction and SR. Feng *et al*. [6] proposed a multi-stage feature fusion mechanism where features from previous stages guide the learning of subsequent target features in multi-contrast SR tasks. Li *et al*. [25] pioneered the application of transformers in multi-contrast MRI SR, introducing a transformer-empowered multi-scale contextual matching and aggregation network. Li *et al*. [27] later enhanced this approach by integrating wavelet transforms with a cross-attention mechanism, further improving performance. However, existing multi-modal representation

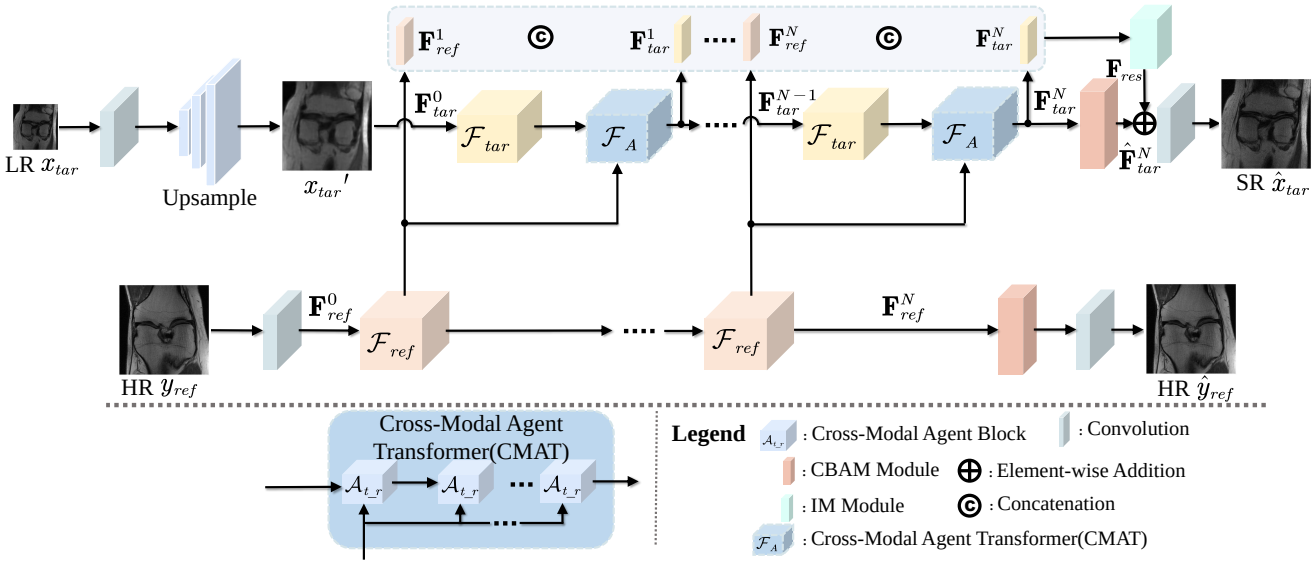


Figure 2. Architecture of the proposed agent-conditioned multi-contrast MRI SR network and cross-modal agent transformer. 'CBAM' and 'IM' refer to the channel-spatial attention module and the layer attention module, respectively, and have the same design as in [6].

learning for MR images SR generally integrates data from different modalities acquired simultaneously from the same subject. In contrast, our method implicitly learns commonalities among multi-contrasts from different subjects.

2.3. Vision Transformer

Transformer was originally proposed as a sequence-to-sequence model in natural language processing (NLP) for tasks such as machine translation [24]. Due to its powerful and flexible modeling capabilities, researchers began exploring its application to computer vision tasks. For example, Dosovitskiy *et al*. [5] first proposed the Vision Transformer (ViT), which splits an image into multiple fixed-size patches and processes these patches as sequences. Following this, various variants and improvements of the transformer framework emerged [15]. ViT's powerful learning capability benefits from the self-attention mechanism. However, it faces challenges due to the quadratic complexity of Softmax attention. To reduce computational costs, several variants have been proposed, such as the sparse global attention of PVT [42], the convolution-like attention of NAT [17], and the deformable attention of DAT [45]. These methods, however, inherently limit the global receptive field of self-attention. On the other hand, linear attention addresses this issue by reducing the complexity to $\mathcal{O}(N)$ [21]. For instance, FLatten Transformer [13] introduced a focused function and adopted depthwise convolution to maintain feature diversity. Efficient Attention [38] applied the Softmax function to both Q and K. While these methods are effective, the expressive power of linear attention remains limited.

In this work, we propose a high expressiveness and efficient cross-modal agent transformer, where the agent is trained to capture commonalities from the auxiliary contrast and guide the target contrast to dynamically adjust and learn these commonalities, thus achieving SR tasks between unpaired MR data.

3. Methodology

3.1. Overall Architecture

Our approach adopts a novel perspective, allowing the use of an HR reference image $y_{ref} \in \mathbb{R}^{H\times W}$ from different subjects to guide the SR reconstruction of the target image $x_{tar} \in \mathbb{R}^{h\times w}$. Specifically, for a $2\times$ enlargement scale, the HR reference image y_{ref} has dimensions $H = W = 320$, while the target image x_{tar} has input dimensions $h = w = 160$. As illustrated in Fig. 2, our proposed network can accept HR reference images from any patient as an auxiliary input. We explore the commonalities and distinctiveness of the same stage features from the two branches. Subsequently, we capture agent commonalities that are not specific to any contrast. We then broadcast these commonalities back to each target feature, allowing them to interact with the target contrast. This process dynamically adjusts and learns to adapt to the commonalities of the target contrast. This approach better transfers knowledge between unpaired data and provides beneficial features for unpaired target images and subjects.

3.1.1 Feature Iterative Extraction

To explore commonalities and distinctiveness from auxiliary contrast images of different subjects, we use a cascaded residual layer [50] to extract multi-stage features. First, we apply separate 3×3 convolutional layers to extract the initial LR and HR representations. Specifically, this operation converts the single-channel input to a 64-channel feature representation, where the HR reference representation is denoted as $\mathbf{F}_{ref}^{0} \in \mathbb{R}^{64\times320\times320}$. Here, we rescale the LR target representation of size $64 \times 160 \times 160$, according to the scaling factor of the SR task to match the spatial dimensions of $\mathbf{F}_{ref}^{0}$. The feature representation of each branch at any stage is as follows:

$$\mathbf{F}_{tar}^{n+1} = \mathcal{F}_A(\mathbf{F}_{tar}^{n}, \mathbf{F}_{ref}^{n}), \quad \mathbf{F}_{ref}^{n} = \mathcal{F}_{ref}^{n}(\mathbf{F}_{ref}^{n-1}). \quad (1)$$

Here, n and N ($n \in N$) represent the index of the current residual group and the total number of residual groups across all branches, respectively. $\mathcal{F}_A$ denotes the cross-modal agent transformer (CMAT) module, and $\mathcal{F}_{ref}^{n}$ is the residual group at the n^{th} stage. Throughout the entire process, both residual features $\mathbf{F}_{tar}^{n}$ and $\mathbf{F}_{ref}^{n}$ maintain a consistent feature dimensionality of $64 \times 320 \times 320$, ensuring spatial alignment across all residual groups and promoting effective knowledge transfer across different contrasts and subjects.

3.1.2 Agent-Conditioned Feature Learning

Inspired by [14], we designed the CMAT module to identify the inherent connections between reference and target contrasts, even in the absence of direct pairing. Specifically, we use agent tokens as intermediaries for transferring key knowledge across contrasts and between subjects. By expanding the traditional attention mechanism from a simple triplet (Query, Key, Value) to a quadruplet that includes an ***Agent*** component, the agent token A can capture commonalities and specific knowledge. As shown in Fig. 2, the shallow features $\mathbf{F}_{tar} \in \mathbb{R}^{64\times320\times320}$ and $\mathbf{F}_{ref} \in \mathbb{R}^{64\times320\times320}$ are fed into the CMAT module to extract and capture shareable foundational anatomical and tissue structure features from auxiliary contrasts. Our specifically designed agent tokens A originate from reference contrast data to guide the model to concentrate on task-specific features. To further facilitate the interaction between the agent and the target contrast, our designed query vector Q originates from the target branch. Formally, we have the following definitions:

$$\begin{aligned}(Q, A, K, V) =&(\mathtt{W}\mathbf{F}_{tar}, \mathtt{avgpool}(\mathtt{W}\mathbf{F}_{ref}),\\ &\mathtt{W}\mathbf{F}_{ref}, \mathtt{W}\mathbf{F}_{ref}),\end{aligned} \quad (2)$$

where $\mathtt{W}$ denotes the matrix for generating tokens, and $\mathtt{avgpool}$ is an average pooling operation. The query vector Q formulated as $Q \in \mathbb{R}^{H\times L\times d}$, where $H = 8$ is the number of attention heads, $L = 102400$ is the number of tokens, and $d = 8$ is the per-head embedding dimension. Similarly, the key vector K and value vector V have the same shape as Q, i.e., $K, V \in \mathbb{R}^{H\times L\times d}$. The agent tokens A are defined as $A \in \mathbb{R}^{H\times L_a\times d}$, where $L_a = 49$ is the number of agent tokens.

Through this agent-conditioned mechanism, knowledge transfer across different contrasts and subjects is achieved. These agent tokens not only learn commonalities but also effectively broadcast this information back to the query tokens Q of the target contrast, thereby dynamically adjusting cross-subject reference contrast data and supplementing beneficial features. The above process is formulated as follows:

$$\begin{aligned}\mathbf{F}_{tar}^{n+1} =&\mathtt{Reshape}(\mathtt{Softmax}(Q, A, \mathtt{Softmax}(A, K, V))\\ &+ \mathtt{DWC}(V)) + \mathbf{F}_{tar}^{n},\end{aligned} \quad (3)$$

where $\mathtt{Softmax}$ denotes Softmax attention, and $\mathtt{DWC}$ denotes a 3×3 depthwise convolutional operation. It is worth noting that our designed CMAT contains multiple cross-modal agent blocks. In these blocks, the tokens for K and V are computed from the feature $\mathbf{F}_{ref}$, while the tokens for Q are computed from the output of the previous block. This design allows the agent tokens to continuously transfer knowledge of commonalities and distinctiveness in data across subjects, enabling the target branch can absorb foundational anatomical details.

3.1.3 Image Reconstruction

Finally, inspired by [6], we employ a CBAM and an IM module to reveal responses from all dimensions of the feature map. The output of the final CMAT, $\mathbf{F}_{tar}^{N} \in \mathbb{R}^{64\times320\times320}$, is fed into the CBAM module to obtain the $\hat{\mathbf{F}}_{tar}^{N} \in \mathbb{R}^{64\times320\times320}$. Furthermore, our model stores the intermediate features at each stage. To maintain feature diversity, these features are activated through the linear layer IM module to obtain the enriched representation $\mathbf{F}_{res} \in \mathbb{R}^{64\times320\times320}$. After that, a 1×1 convolutional layer is used to obtain the final reconstructed target SR image $\hat{x}_{tar} \in \mathbb{R}^{1\times320\times320}$, which can be written as:

$$\hat{x}_{tar} = \mathtt{Conv}(\mathbf{F}_{res} \oplus \hat{\mathbf{F}}_{tar}^{N} \oplus \mathbf{F}_{tar}^{0}), \quad (4)$$

where $\oplus$ means element-wise summation.

3.1.4 Loss Function

The L_1 loss is used to evaluate the SR results of the target image:

$$\mathcal{L} = \lambda_{tar}\|\hat{x}_{tar} - x_{tar}\prime\|_1 + \lambda_{ref}\|\hat{y}_{ref} - y_{ref}\|_1, \quad (5)$$

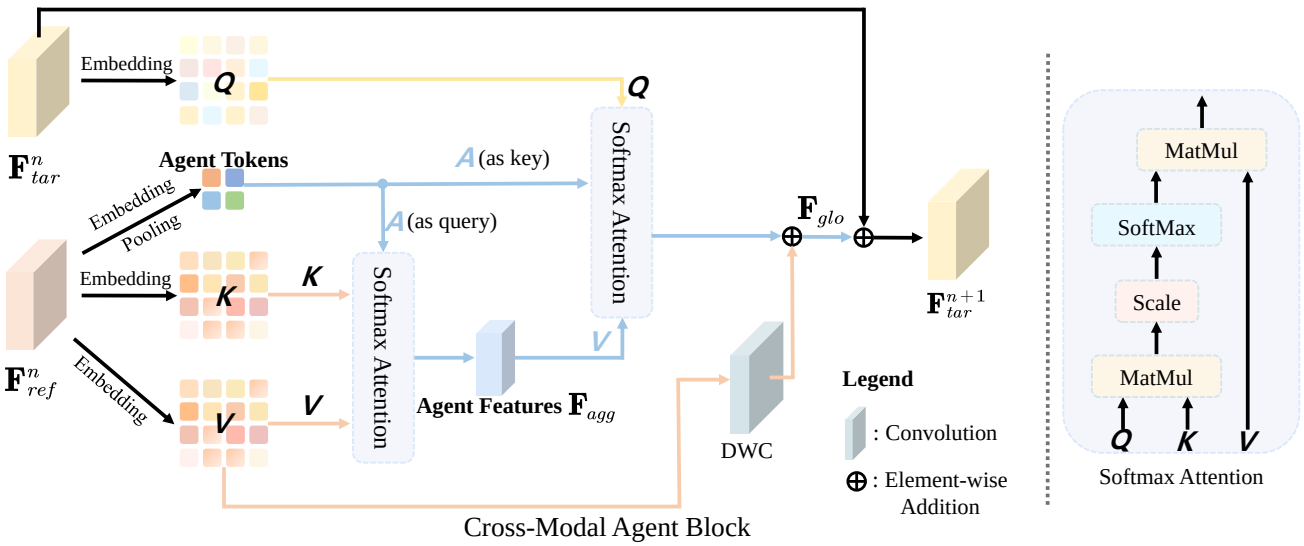


Figure 3. Architecture of the proposed Cross-Modal Agent Block, where the right side of the dashed line represents the computation process of Softmax attention.

where λ_{tar} and λ_{ref} weigh the trade-off between the target image and reference image reconstruction.

3.2. Cross-Modal Agent Block

As shown in Fig. 3, the agent tokens $A \in \mathbb{R}^{H\times L_a\times d}$ act as intermediaries for $Q \in \mathbb{R}^{H\times L\times d}$ and $K \in \mathbb{R}^{H\times L\times d}$, collecting commonalities knowledge from the reference branch and then broadcasting these commonalities back to the target branch. Recall that Eq. 2 provides an overview of the representations of query, key, value, and agent tokens. Agent tokens A, serving as a commonalities feature query agent, capture commonalities from reference contrasts during training. These commonalities reflect shareable foundational anatomical and tissue structure features rather than specific details of a particular contrast. Therefore, the agent features we capture $\mathbf{F}_{agg} \in \mathbb{R}^{H\times L_a\times d}$ are characterized by foundational anatomical commonalities. Furthermore, the introduced agent bias helps maintain spatial consistency in the model. The features of agent commonalities $\mathbf{F}_{agg}$ can be represented as:

$$\mathbf{F}_{agg} = \texttt{Softmax}\left(\frac{A(K)^\top}{\sqrt{d_k}} + \mathbf{B}_1\right)V, \tag{6}$$

where $\mathbf{B}_1 \in \mathbb{R}^{H\times L_a\times L}$ is the agent bias for the commonalities calculation, $d_k = 8$ represents the channel dimension of K, and the calculation methods for A, K, Q and V refer to Eq. 2. After that, we use A as the key (differently from the first instance) and the features of agent commonalities $\mathbf{F}_{agg}$ as the value, while employing the original query matrix Q for a second round of global attention calculation. This process enables the precise allocation and broadcasting of the features of agent commonalities onto the query tokens of the target contrast. Consequently, this approach guides the target contrast to focus on learning commonalities while enhancing its understanding of the distinctiveness of different anatomical structures. This method delves deeper into the intrinsic connections between the auxiliary and target contrasts, even though they are unpaired. To augment the diversity of target features, a depthwise convolution is employed at the end to preserve the feature diversity of the reference branch, and can be expressed as:

$$\mathbf{F}_{glo} = \texttt{Softmax}\left(\frac{Q(A)^\top}{\sqrt{d_k}} + \mathbf{B}_2\right)\mathbf{F}^V_{agg} + \texttt{DWC}(V), \tag{7}$$

where $\mathbf{B}_2 \in \mathbb{R}^{H\times L\times L_a}$ serves as the agent bias for the second attention calculation, and $\mathbf{F}^V_{agg} \in \mathbb{R}^{H\times L_a\times d}$ represents the value in the second attention calculation is directly computed from features of agent commonalities $\mathbf{F}_{agg}$. Finally, we derive the output of the cross-modal agent block, denoted as $\mathbf{F}_{glo} \in \mathbb{R}^{1\times 102400\times 64}$.

4. Experiments

4.1. Datasets

We evaluated the performance of our proposed network on three datasets, including two in-house brain datasets: **SMS** and **uMR**, and one public multi-contrast MRI dataset: **fastMRI** [46]. For **fastMRI**, the largest open-access raw MR image dataset, we use the unpaired PDWI contrast to guide the SR of FS-PDWI contrast, randomly filter out 200 and 40 PDWI and FS-PDWI brain volumes for training and validation, respectively, as unpaired experimental configurations. The **SMS** dataset was acquired with fully

Table 1. Quantitative metrics results on three datasets with 2× enlargement scales. The best quantitative metrics results are marked in red.

Dataset	fastMRI [46]			SMS			uMR		
Metrics	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓
Unet [36]	0.2287	22.5696	0.2246	0.7678	34.8329	0.0111	0.6515	33.8070	0.0264
UnetSN [43]	0.5034	25.1499	0.1252	0.9754	39.4558	0.0038	0.9813	39.6306	0.0067
McMRSR [25]	0.6767	30.0768	0.0449	0.8969	38.8810	0.0044	0.9709	39.8956	0.0063
MINet [6]	0.6941	31.0488	0.0377	0.9832	41.8756	0.0022	0.9867	42.0970	0.0038
HAT [3]	0.7000	31.4128	0.0356	0.9850	41.7250	0.0023	0.9855	41.3772	0.0045
AgentMRI	0.7007	31.5696	0.0348	0.9867	42.3221	0.0020	0.9904	42.2374	0.0037

Table 2. Quantitative metrics results on three datasets with 4× enlargement scale. The best quantitative metrics results are marked in red.

Dataset	fastMRI [46]			SMS			uMR		
Metrics	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓
Unet [36]	0.0015	12.8584	2.2429	0.5459	27.9188	0.0552	0.4310	27.0451	0.1281
UnetSN [43]	0.0065	14.2090	1.5879	0.8683	31.3687	0.0248	0.8192	32.0166	0.0394
McMRSR [25]	0.5513	27.2725	0.0810	0.4674	31.0003	0.0269	0.8533	33.6797	0.0264
MINet [6]	0.5837	28.4560	0.0633	0.8719	32.8143	0.0175	0.8393	34.2608	0.0230
HAT [3]	0.6004	29.4626	0.0516	0.9198	33.4658	0.0151	0.9413	34.5221	0.0217
AgentMRI	0.6008	29.5927	0.0507	0.9045	33.5515	0.0148	0.9401	34.8800	0.0200

k-space sampling using a clinical 3T Siemens Magnetom Skyra scanner on 155 subjects, where each MR scanning parameter was as follows: T1WI with TR = 2001 ms, TE = 10.72 ms; T2WI with TR = 4511 ms, TE = 112.86 ms. Both sequences had a slice thickness of 5 mm, a matrix size of 320 x 320 x 20, and a field of view of 230 x 200 mm^2. The **uMR** brain dataset was acquired using a 3T whole-body scanner on 50 subjects, where each MR scanning parameter was as follows: T1WI with TR = 2001 ms, TE = 10.72 ms; T2WI with TR = 4511 ms, TE = 112.86 ms. Both sequences had a slice thickness of 4 mm, a matrix size of 320 x 320 x 20, and a field of view of 220 x 250 mm^2. The **SMS** and **uMR** datasets are randomly matched subjects-wise with a ratio of 7:1:2 for the training, validation, and test sets as unpaired experimental configurations.

4.2. Baselines

We compared our AgentMRI with various recent state-of-the-art methods to demonstrate its effectiveness, including two single-contrast SR methods: (1) Unet [36], a widely used convolutional network for biomedical image tasks; (2) UnetSN [43], an SR algorithm that restores low-resolution images with unknown and complex degradation using spectral normalization; (3) HAT [3], an SR algorithm that uses hybrid attention to fully exploit the potential of transformers in image restoration tasks, and two multi-contrast methods: (4) McMRSR [25], an MRI SR algorithm that matches and aggregates multi-scale context between contrasts to model long-range dependencies in both reference and target images; (5) MINet [6], an MRI SR algorithm that integrates multi-stage representations between the reference and target contrasts. For a fair comparison, all baselines were retrained with their predefined parameter configurations.

4.3. Implementation Details.

Our model is implemented in PyTorch and runs on a single NVIDIA RTX 3090 GPU with 24GB of memory. We use the Adam optimizer with a learning rate of 1e-3 and a mini-batch size of 8 for network training over 35 epochs, with the first-moment and second-moment coefficients set to 0.9 and 0.999, respectively. The hyperparameters λ_{tar} and λ_{ref} are set to 0.7 and 0.3, respectively, and N is set to 6. The cross-modal agent attention head number is set to 8, and the patch size is 1×1. Following [14], the number of agent tokens is 49.

4.4. Quantitative Results.

We evaluated our SR results by computing the SSIM, PSNR, and NMSE between the super-resolved image and the fully sampled ground truth image. Tabs. 1 and 2 present

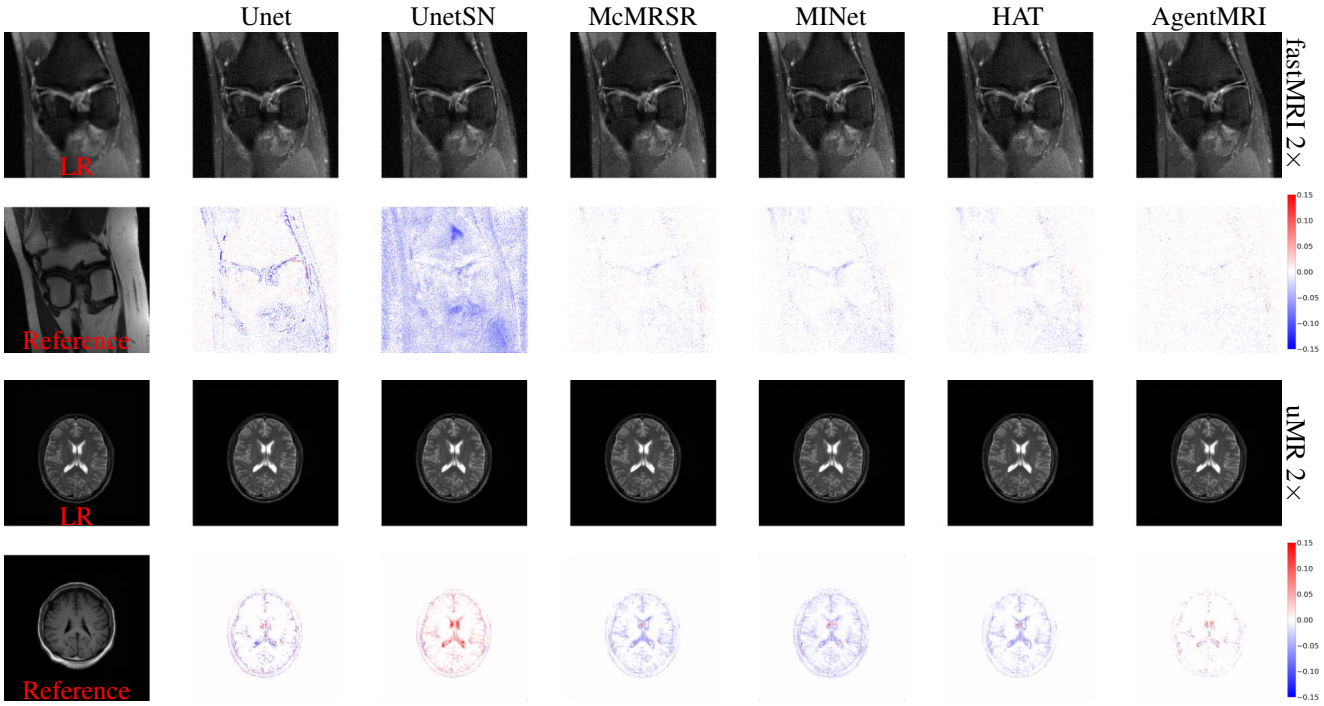


Figure 4. Visual SR results and error maps of different methods on the **fastMRI** and **uMR** datasets with 2× enlargement scale.

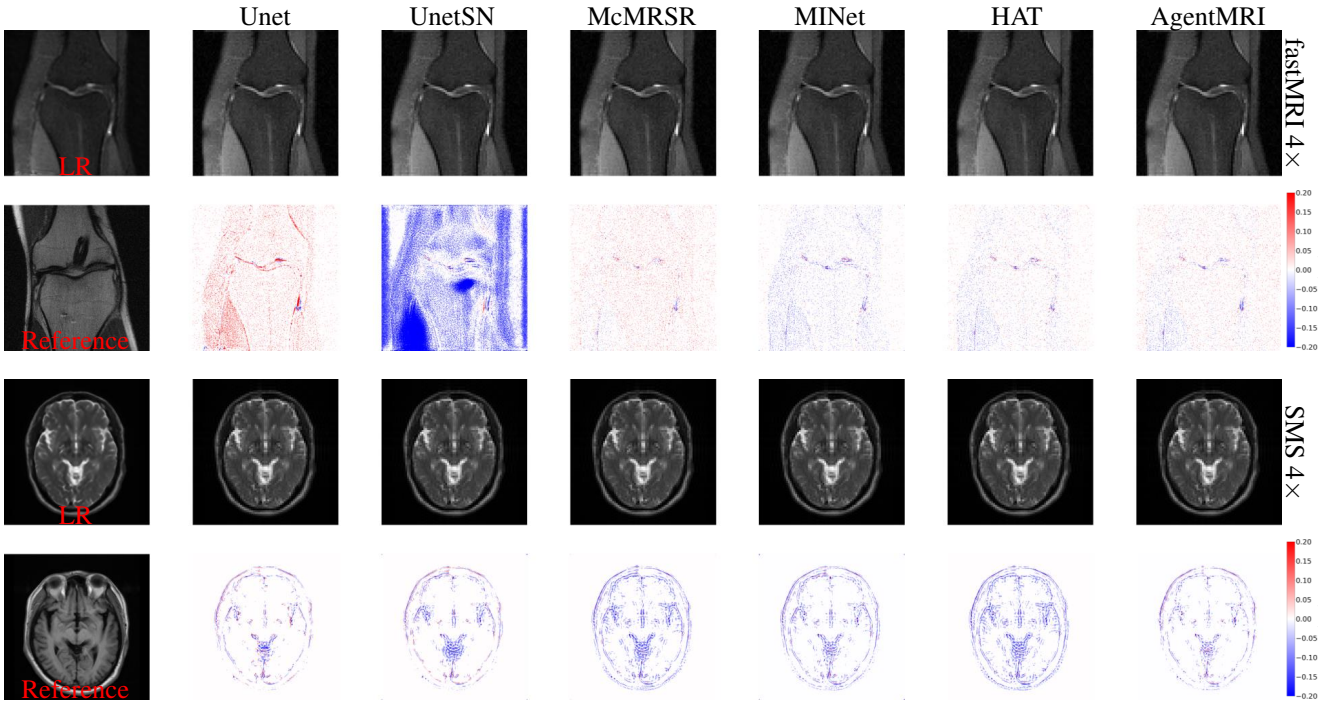


Figure 5. Visual SR results and error maps of different methods on the **fastMRI** and **SMS** datasets with 4× enlargement scale.

the quantitative comparison between our proposed method and various baselines under 2× and 4× enlargement using unpaired configurations, respectively. As shown, our method yields the best results in terms of PSNR and NMSE metrics, demonstrating that AgentMRI can effectively explore the intrinsic connection between the reference contrast and the target contrast, even when they are not directly paired. We note that single-contrast methods are

far less effective than multi-contrast models. Furthermore, multi-contrast SR models are less effective than our method because they struggle to capture foundational anatomical features from unpaired data. These models are unable to achieve effective interaction between reference and target contrasts, resulting in the inability of the captured commonalities and distinctiveness knowledge to adapt to target contrast images effectively. More importantly, even though reconstructing SR images at 4× enlargement, our method can still outperform previous methods, which can be attributed to providing beneficial information supplements for target contrast data across subjects through an agent-conditioned mechanism. Although SSIM did not significantly outperform existing methods in some experiments, we attribute this to the sensitivity of SSIM to noise in the dataset. In contrast, the improvements in PSNR and NMSE more objectively reflect the robustness of the model. Our AgentMRI achieves 31.5696 dB and 29.5927 dB in PSNR on the **fastMRI** dataset, 42.3221 dB and 33.5515 dB on the **SMS** dataset, as well as 42.2374 dB and 34.8800 dB on the **uMR** dataset.

4.5. Qualitative Evaluation

To further evaluate the robustness of our method, we conducted a quantitative analysis of the performance of AgentMRI. Figs. 4 and 5 show SR results and error maps for unpaired **fastMRI**, **SMS**, and **uMR** at both 2× and 4× enlargements. The SR images indicate that single-contrast methods can restore the basic structure of the MR image. However, multi-contrast SR methods improve the results, with fewer structural losses. Specifically, our AgentMRI produces high-quality images with clear details, minimal checkerboard effects, and less structural loss, effectively restoring the entire structure of the knee or brain. This is attributed to the proposed agent-conditioned mechanism, which has excellent learning abilities in both commonalities and distinctiveness. More importantly, the error maps for different enlargement scales demonstrate that our method yields the smallest errors across various datasets.

4.6. Ablation Study

4.6.1 Component Analysis

Here, we investigate the importance of each key component of AgentMRI. To verify whether the agent-conditioned multi-contrast MRI SR possesses cross-subject adaptability, we designed seven different variants, i.e., $w/o\ \mathbf{B}_1$, which uses cross-modal Softmax attention calculation [40] without capturing the spatial information of agent commonalities through agent bias; $w/o\ \mathbf{B}_2$, which directly employ agent feature $\mathbf{F}_{agg}$ for the second Softmax attention calculation to evaluate whether agent bias can direct focus to task-specific features by sharing agent commonalities with the target branch; w/o `DWC`, which uses agent attention without a depthwise convolution (`DWC`) module to evaluate whether the knowledge of commonalities and distinctiveness from the unpaired reference and target contrast can inject diversity features into the target image; $w\ \mathbf{B}_1$, which performs only Softmax attention calculation between the agent bias $\mathbf{B}_1$ and agent tokens A; $w\ \mathbf{B}_2$, which performs only Softmax attention calculation between the agent features $\mathbf{F}_{agg}$ and agent tokens A; w `DWC`, which only uses a depthwise convolution (`DWC`) module without agent attention. We conducted additional ablation studies by restoring and comparing the vanilla attention with our CMAT, referred to as w `Cross-Vanilla Att.`. The vanilla attention follows the traditional Query-Key-Value triplet, where the query vector is generated from target image features, and the key and value vectors are derived from the reference contrast data. Unlike our CMAT, vanilla attention does not use agent features to extract cross-subject shared information, relying instead on reference contrast data for feature fusion. The quantitative metrics for each variant model across three datasets at the 2× enlargement scale and the 4× enlargement scale are presented in Tabs. 3 and 4, respectively. The qualitative results are also presented in Fig. 6. It can be observed that all variant models perform worse than our AgentMRI. This indicates that removing these key components leads to a performance decline, thereby verifying their importance in AgentMRI. Specifically, $w/o\ \mathbf{B}_1$ has the lowest performance, supporting our initial hypothesis that even unpaired multi-contrast data have commonalities that can guide the SR of the target contrast. $w/o\ \mathbf{B}_2$ outperforms $w/o\ \mathbf{B}_1$ because the agent bias can better guide the reference contrast to adapt to the target image. Similarly, the performance of w/o `DWC` also shows a decline, as the depthwise convolution module helps to inject diversity features. Furthermore, the experimental results show that when only $\mathbf{B}_1$, $\mathbf{B}_2$, and `DWC` are used, the performance is inferior to that of the complete AgentMRI model. w `Cross-Vanilla Att.` underperforms compared to AgentMRI because the model fails to integrate cross-subject auxiliary contrast data into the target image without agent tokens, as evidenced by the lower PSNR and SSIM metrics, and the higher NMSE metric. In summary, AgentMRI outperforms other models, demonstrating its robust ability to maximize the use of commonalities in unpaired data.

4.6.2 Hyperparameter Analysis

We conducted a hyperparameter ablation analysis on the **uMR** dataset to evaluate the impact of different parameter settings for the weights λ_{tar} and λ_{ref}. As shown in Tab. 5, the results indicate that the trade-off between the values of λ_{tar} and λ_{ref} is important for optimal model performance. Specifically, when the weights are equal, the model

Table 3. Ablation study on different variants under three datasets with a 2× enlargement scale. The best quantitative metrics results are marked in red.

Dataset	fastMRI [46]			SMS			uMR		
Metrics	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓
$w/o\ \mathbf{B}_1$	0.6925	30.9664	0.0384	0.9729	41.9054	0.0022	0.9666	41.8241	0.0041
$w/o\ \mathbf{B}_2$	0.6843	30.6074	0.2246	0.9726	42.2453	0.0020	0.9395	40.8532	0.0051
w/o DWC	0.6906	30.8367	0.0388	0.9779	41.2401	0.0025	0.9664	41.7416	0.0041
w Cross-Vanilla Att.	0.6862	30.7800	0.0395	0.9768	41.0214	0.0027	0.9871	42.0681	0.0038
$w\ \mathbf{B}_1$	0.6973	31.2315	0.0362	0.9851	42.0130	0.0027	0.9833	42.0570	0.0038
$w\ \mathbf{B}_2$	0.6937	30.9651	0.0381	0.9841	42.1152	0.0021	0.9738	41.9699	0.0039
w DWC	0.6980	31.3462	0.0358	0.9860	42.1825	0.0020	0.9835	42.2111	0.0037
AgentMRI	0.7007	31.5696	0.0348	0.9867	42.3221	0.0020	0.9904	42.2374	0.0037

Table 4. Ablation study on different variants under three datasets with a 4× enlargement scale. The best quantitative metrics results are marked in red.

Dataset	fastMRI [46]			SMS			uMR		
Metrics	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓
$w/o\ \mathbf{B}_1$	0.4604	25.3290	0.1307	0.8324	33.0821	0.0165	0.8494	33.6912	0.0268
$w/o\ \mathbf{B}_2$	0.5323	26.7480	0.0924	0.7589	33.0083	0.0168	0.7789	33.4524	0.0285
w/o DWC	0.5356	26.7596	0.0940	0.8439	33.2791	0.0157	0.8196	32.7649	0.0338
w Cross-Vanilla Att.	0.5485	27.3697	0.0815	0.7949	30.4481	0.0315	0.9158	34.3611	0.0226
$w\ \mathbf{B}_1$	0.5768	27.7572	0.0726	0.9015	33.3044	0.0157	0.9090	34.5627	0.0215
$w\ \mathbf{B}_2$	0.5430	26.813	0.0907	0.8277	32.7762	0.0180	0.8323	32.6149	0.0354
w DWC	0.5873	28.6027	0.0600	0.9019	33.3554	0.0155	0.9029	34.6077	0.0212
AgentMRI	0.6008	29.5927	0.0507	0.9045	33.5515	0.0148	0.9401	34.8800	0.0200

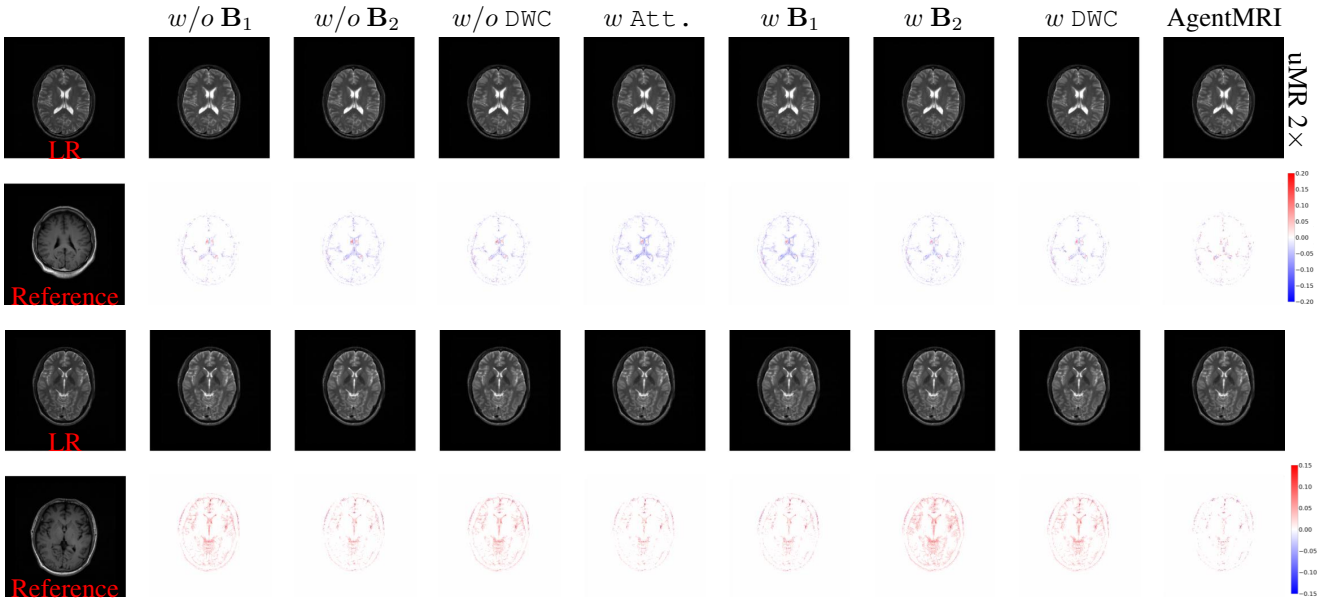


Figure 6. Ablation study of the key components in our method, where $w/o\ \mathbf{B}_1$, $w/o\ \mathbf{B}_2$, w/o DWC, w Att., $w\ \mathbf{B}_1$, $w\ \mathbf{B}_2$, and w DWC represent seven variations of our model. Here, w Att. refers to w Cross-Vanilla Att.. Visual SR results and error maps of different variations on the **uMR** dataset with 2× enlargement scale.

Table 5. Ablation study on different parameter settings under **uMR** with a 2× enlargement scale. The best quantitative metrics results are marked in red.

Hyperparameter	SSIM↑	PSNR↑	NMSE↓
λ_{tar}=0.5, λ_{ref}=0.5	0.9861	42.2273	0.0037
λ_{tar}=0.3, λ_{ref}=0.7	0.9680	41.3301	0.0046
λ_{tar}=0.8, λ_{ref}=0.2	0.9681	41.8269	0.0040
λ_{tar}=0.2, λ_{ref}=0.8	0.9621	40.8509	0.0051
λ_{tar}=0.5, λ_{ref}=1	0.9836	42.2365	0.0037
λ_{tar}=1, λ_{ref}=0.5	0.9847	42.1555	0.0037
λ_{tar}**=0.7,** λ_{ref}**=0.3**	0.9904	42.2374	0.0037

shows the second-best performance. However, when the model relies too heavily on reference contrast, performance significantly declines, suggesting that reference data alone cannot adequately restore the fine details of the target image. Similarly, an overemphasis on target contrast does not result in performance improvements, demonstrating the indispensable role of unpaired reference contrast in providing task-specific information and supplementary details. The best performance was observed when the weights slightly favored the target contrast, indicating that prioritizing the target data helps enhance SR quality, while the unpaired reference data still provides beneficial features. Based on the results of the ablation study, we find that setting λ_{tar} to 0.7 and λ_{ref} to 0.3 yields the best performance.

5. Conclusion

In this work, we focus on exploring the commonalities and distinctiveness between unpaired MR images to enhance MRI with target contrast. For this purpose, we introduce agent-conditioned multi-contrast MRI SR with cross-subject adaptation, which can achieve SR of target images under the guidance of any HR reference contrast. Specifically, AgentMRI mines commonalities independent of the target contrast through a trained agent and then interacts with the target contrast to guide the model to focus on task-specific features. In the future, we will explore repurposing the abundant yet untapped medical image resources to further investigate the potential relationships between unpaired data.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgement

This work was supported in part by the CCF-ZhiPu Large Model Fund Project (202212), in part by the Textile's Light Applied Basic Research Project (J202209), in part by The Hong Kong Polytechnic University (PolyU) under Grant P0042740, and in part by the PolyU Research Institute for Sports Science and Technology under Grant P0044571.

References

[1] M. Akçakaya, S. Moeller, S. Weingärtner, and K. Uğurbil. Scan-specific robust artificial-neural-networks for k-space interpolation (raki) reconstruction: database-free deep learning for fast imaging. *Magnetic resonance in medicine*, 81(1):439–453, 2019. 2

[2] B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento. Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, 11(4):900–915, 2021. 2

[3] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023. 7

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[5] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[6] C.-M. Feng, H. Fu, S. Yuan, and Y. Xu. Multi-contrast mri super-resolution via a multi-stage integration network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 140–149. Springer, 2021. 2, 3, 4, 5, 7

[7] C.-M. Feng, B. Li, X. Xu, Y. Liu, H. Fu, and W. Zuo. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8064–8073, June 2023. 2

[8] C.-M. Feng, Y. Yan, G. Chen, Y. Xu, Y. Hu, L. Shao, and H. Fu. Multimodal transformer for accelerated mr imaging. *IEEE Transactions on Medical Imaging*, 42(10):2804–2816, 2023. 2

[9] C.-M. Feng, Y. Yan, H. Fu, L. Chen, and Y. Xu. Task transformer network for joint mri reconstruction and super-resolution. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 307–317. Springer, 2021. 3

[10] S. E. Gentry, E. Chow, N. Dzebisashvili, M. Schnitzler, K. Lentine, C. Wickliffe, E. Shteyn, J. Pyke, A. Israni, B. Kasiske, et al. The impact of redistricting proposals on health care expenditures for liver transplant candidates and recipients. *American Journal of Transplantation*, 16(2):583–593, 2016. 2

[11] G. H. Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2):133–139, 2011. 1

[12] J. P. Haldar, D. Hernando, and Z.-P. Liang. Compressed-sensing mri with random encoding. *IEEE transactions on Medical Imaging*, 30(4):893–903, 2010. 3

[13] D. Han, X. Pan, Y. Han, S. Song, and G. Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5961–5971, October 2023. 4

[14] D. Han, T. Ye, Y. Han, Z. Xia, S. Song, and G. Huang. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2312.08874*, 2023. 5, 7

[15] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 4

[16] R. Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007. 3

[17] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, June 2023. 4

[18] M. Jiang, M. Zhi, L. Wei, X. Yang, J. Zhang, Y. Li, P. Wang, J. Huang, and G. Yang. Fa-gan: Fused attentive generative adversarial networks for mri image super-resolution. *Computerized Medical Imaging and Graphics*, 92:101969, 2021. 3

[19] Q. Jiang, C. Chen, H. Zhao, L. Chen, Q. Ping, S. D. Tran, Y. Xu, B. Zeng, and T. Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7661–7671, June 2023. 3

[20] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE transactions on image processing*, 26(9):4509–4522, 2017. 3

[21] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 4

[22] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*, 2022. 3

[23] P. Lei, F. Fang, G. Zhang, and T. Zeng. Decomposition-based variational network for multi-contrast mri super-resolution and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21296–21306, October 2023. 2

[24] M. Lewis. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 4

[25] G. Li, J. Lv, Y. Tian, Q. Dou, C. Wang, C. Xu, and J. Qin. Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20636–20645, June 2022. 2, 3, 7

[26] G. Li, J. Lv, X. Tong, C. Wang, and G. Yang. High-resolution pelvic mri reconstruction using a generative adversarial network with attention and cyclic loss. *IEEE Access*, 9:105951–105964, 2021. 3

[27] G. Li, J. Lyu, C. Wang, Q. Dou, and J. Qin. Wavtrans: Synergizing wavelet and cross-attention transformer for multi-contrast mri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 463–473. Springer, 2022. 2, 3

[28] G. Li, C. Rao, J. Mo, Z. Zhang, W. Xing, and L. Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11365–11374, June 2024. 2

[29] Y. Liu, L. Fan, C. Zhang, T. Zhou, Z. Xiao, L. Geng, and D. Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021. 3

[30] Q. Lyu, H. Shan, C. Steber, C. Helis, C. Whitlow, M. Chan, and G. Wang. Multi-contrast super-resolution mri through a progressive network. *IEEE Transactions on Medical Imaging*, 39(9):2738–2749, 2020. 2, 3

[31] J. McGinnis, S. Shit, H. B. Li, V. Sideri-Lampretsa, R. Graf, M. Dannecker, J. Pan, N. Stolt-Ansó, M. Mühlau, J. S. Kirschke, et al. Single-subject multi-contrast mri super-resolution via implicit neural representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–183. Springer, 2023. 2

[32] E. Plenge, D. H. Poot, M. Bernsen, G. Kotek, G. Houston, P. Wielopolski, L. van der Weerd, W. J. Niessen, and E. Meijering. Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magnetic resonance in medicine*, 68(6):1983–1993, 2012. 1

[33] D. Qiu, Y. Cheng, and X. Wang. Gradual back-projection residual attention network for magnetic resonance image super-resolution. *Computer Methods and Programs in Biomedicine*, 208:106252, 2021. 3

[34] D. Qiu, S. Zhang, Y. Liu, J. Zhu, and L. Zheng. Super-resolution reconstruction of knee magnetic resonance imag-

ing based on deep learning. *Computer methods and programs in biomedicine*, 187:105059, 2020. 3

[35] X. Qu, Y. Hou, F. Lam, D. Guo, J. Zhong, and Z. Chen. Magnetic resonance image reconstruction from undersampled measurements using a patch-based nonlocal operator. *Medical image analysis*, 18(6):843–856, 2014. 2

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 7

[37] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge. Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sensing*, 12(14):2207, 2020. 3

[38] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3531–3539, January 2021. 4

[39] S. Tourbier, X. Bresson, P. Hagmann, J.-P. Thiran, R. Meuli, and M. B. Cuadra. An efficient total variation algorithm for super-resolution in fetal brain mri with adaptive regularization. *NeuroImage*, 118:584–597, 2015. 3

[40] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 3, 9

[41] E. Van Reeth, I. W. Tham, C. H. Tan, and C. L. Poh. Super-resolution in magnetic resonance imaging: a review. *Concepts in Magnetic Resonance Part A*, 40(6):306–325, 2012. 1

[42] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, October 2021. 4

[43] X. Wang, L. Xie, C. Dong, and Y. Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, October 2021. 7

[44] Z. Wang, Z. Cui, and Y. Zhu. Multi-modal medical image fusion by laplacian pyramid and adaptive sparse representation. *Computers in Biology and Medicine*, 123:103823, 2020. 2

[45] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4803, June 2022. 4

[46] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018. 6, 7, 10

[47] C. Zhang, Z. Yang, X. He, and L. Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. 3

[48] Y. Zhang, X. He, Y. H. Chan, Q. Teng, and J. C. Rajapakse. Multi-modal graph neural network for early diagnosis of alzheimer's disease from smri and pet scans. *Computers in Biology and Medicine*, 164:107328, 2023. 3

[49] Y. Zhang, K. Li, K. Li, and Y. Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13425–13434, June 2021. 3

[50] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 5

[51] B. Zhou and S. K. Zhou. Dudornet: Learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2